On Linguistic Formulas for Knowledge Representation

Inger Bierschenk

1983

Gruppen för kognitionsvetenskaplig forskning,
Lunds universitet

Linguistics has gained very much from the formal sciences, especially in computer age. Today, artificial intelligence (AI) and other fields of computer science are working with linguistically founded models in increasing degree. It seems, however, as if the linguistic theories are beginning to be historic. At least they are less visible in the literature, whereas language understanding models, as invented by AI, take up substantial place. For example, knowledge representation, an area of rapid growth, which, by necessity, builds on one or several linguistic models (Bobrow & Winograd, 1977), would be studied only halfway, if this basis will not become known. The importance of investigating theories underlying AI applications is further emphasized by the fact that, in the AI community, there can be noticed a tendency to adapt a linguistic model in ways that are not always explicitly described.

This paper takes up one model that has been basic for language-based computer systems. With this background, one primary computer application will be examined. The aim is to illustrate some problems in taking advantage of a theory developed in one field when constructing a linguistic representation in another.

A Formal Construction of Language

In its closest relation to the formal sciences, linguistics works in agreement with the position that human intentional abstracting in observing natural phenomena results in a general symbolic expression. This means that human language, being a natural phenomenon, can only be fully explained through a formal language of description. Thereby natural language is characterized by a definable form. Within the unit of sentence, the rules of a grammar state whether the sequencing of symbols is a true description of a certain language or not. In the development from simple phrase structure grammars to conceptually oriented, the notion of grammaticality, as specified for each grammar, is discussed as being deeply connected with " meaning ". This accentuates the question of what in the language description stands for meaning, that is, the specification and distinction

of the symbol itself.

In a phrase structure grammar (Chomsky, 1957), grammaticality is described in terms of immediate constituency rules, which allow a sentence to be formalized by constituent labels at different syntactic levels. Such a grammar can be looked upon as a set of instructions by which a sentence is " derived ". As an example, the sentence " The block supports the pyramid " can be described and derived as follows:

Applied rules	Derivation
(1) Sentence → NP+VP	(1) NP+VP
(2) NP → T+N	(2) T+N+VP
(3) VP → Verb+NP	(3) T+N+Verb+NP
(4) T → the	(4) the+N+Verb+NP
(5) N → block, pyramid	(5) the+block+Verb+NP
(6) Verb → support	(6) the+block+support+NP
	(2) the+block+support+T+N
	(4) the+block+support+the+N
	(5) the+block+support+the+pyramid

The second level of the derivation is formed by " rewriting " Sentence as NP+VP according to the first rule. This view of language is most commonly displayed in a so called tree structure diagram, which on the one hand does not state the order of the rules but clearly represents the sentence as hierarchic on the other. A redefinition of the main " category " symbols NP and VP into symbols of grammatical function makes it evident that the first NP corresponds to subject and the second to object, and that, consequently, the description presupposes a tighter connection of the verb to the object than to the subject.

The grammar type described is called " context free ", which means that the rules are claimed to be universal. It is obvious from the outcome of this grammar (see last line of derivation) that it does not provide for a correct " spelling " of the language in question unless a grammatical context is considered, that is, a morphological. Variable inflections and suf-

fixes thus have to be added to the grammatical rules making them " context sensitive ". For example

NP (sing) + Verb → NP (sing) + supports (1)

indicates that Verb is rewritten " supports " in the context of a singular NP. This so called generalization of the formalism limits the application of rules to certain grammatical contexts. In syntactic theory, a common language of description and common principles in forming the rules were intended to guarantee that different languages could be described in terms of their common characteristics.

The principle drawback of simple phrase structure grammars for adequate description of language is well-known: declarative statements are highly artificial. Therefore, symbols have been introduced that allow the grammar to describe natural variations on several levels. For example, sentences can be other than declaratives, verbs may indicate mood (the Aux component), and not any word from a vocabulary can be the outcome of an NP generation. Hence, in addition to rewrite rules that apply to category symbols another kind of rewrite rules is supposed to be basic component. These rules are called " selectional " and apply to symbols for lexical categories introducing complex symbols, which are a set of syntactic features (Chomsky, 1965). These features are used as lexical entries in specifying the vocabulary choice. According to this grammar, a noun can be subcategorized in terms of a hierarchic branching

$$N \rightarrow [+\ N,\ \pm\ Common] \qquad (2)$$
$$[+\ Common] \rightarrow [\pm\ Count],\ \text{etc}$$

distinguishing the nouns of the lexicon from each other and inserting them on the right syntactic level. Similarly, the verbs may be described as permitting, for example, Abstract Subject and Animate Object. Chomsky (1965, Chapter 2) exemplifies with

the sentence " Sincerity frightens the boy ", which makes the feature system transparent. In the example sentence " The block supports the pyramid " the verb provides no selection possibility as to the Subject - Object relation between the nouns. These, in their turn, are hardly separable through this kind of syntactic-semantic features being of the same type. Instead, measurable features pertaining to size and shape must specify the difference between them. Through the concept of selection restriction a discussion of " interpretability " has extended the concept of grammaticalness. Pure logical features, however, restrict grammaticalness to pure computation, by which natural language elements only function as associative links.

Typical of context sensitive grammars is the concept of " transformation ". It is connected with " deep structure " such that rules of transformation " operate " on declarative sentences (" kernels ") in a prescribed order before surface generation. Hence, Transformational Grammar. The probably most discussed is the passive transformation, which applies to a symbol sequence of the form NP - Aux - V - NP in that it interchanges the NP's and adds the grammatical markers be + en to Aux:

Description: $NP - Aux - V - NP$ (3)

Change: $X_1 - X_2 - X_3 - X_4 \rightarrow X_4 - X_2 + be + en - X_3 - by + X_1$

In proposing a meta-language for describing the phenomenon of language its constructors assume its meaning to be reflected through the manner in which the symbols combine and operate. The Chomskyan model is one of the strongest in claiming to contribute to the knowledge of language. At least for the last two decades projects have been in progress with the aim to simulate symbolic information processing by humans, where natural language is made the basis for the processing. The pioneer work concerning natural language processing by computer was Ross Quillian's (1968) " Teachable Language Comprehender ". Since he

refers to Chomsky, this section will examine the kind of knowledge represented linguistically in his model of processing.

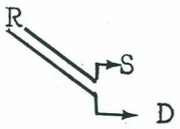
Memory for Processing

The purpose of simulating comprehension necessarily imposes a cognitive aspect on the linguistic model. As a consequence, some processing device is added to the grammar and a mechanism for accumulating the kind of knowledge assumed to be relevant for comprehension and to be extracted from natural language sentences. The notion " memory " in simulations of language comprehension corresponds to deep structure. Therefore, when a linguistic model is consistently employed the meaning of sentences shall be derived from surface variations and be the basis for memory construction and use. Quillian's purpose was to build a memory for conceptual structure and he was in the need of a linguistic model by the help of which deep structure elements can be derived. Quillian (1968) makes the following statement:

> (...) language is remembered, dealt with in thought, and united to non-linguistic concepts in a form that looks like the result of phrase structure rules - what Chomsky calls the 'base phrase marker' or 'basis' of a sentence. (p 222).

In addition to " non-linguistic concepts " it is also talked about " linguistic concepts ", which correspond to words. This means that a syntactic model of the Chomskyan type (Aspects model) will suite the purpose. It could, therefore, be expected an exemplification of how such a grammar applies in the parametrization of sentences.

Quillian presents three parameters, S for the subject of a clause, D for direct object and M for " modificand ", which modifies another word in the same clause (note: the unit is not sentence). S and D are related in a way indicated by a relation, R, which can be formulated with a statement in predicate-logics, R (S, D). Thus a clause gets the following formulation:



(4)

The symbols ( ) represent prepositions and are being substituted with these when prepositions may be utilized as links between verbs and their subjects and objects. Quillian explains that subject of a clause needs not be a subject in the linguistic sense. This turns out to be somewhat obscure, since the symbol for preposition in the formula seems to indicate at least two kinds of linguistic subjects. It would have been clarifying to get the definition of a non-linguistic subject. It would also have been more relevant to get an explanation of the grounds for letting prepositions determine direct objects, because, here, the linguistic sense of " direct " is different from the one employed. In this respect Quillian even leaves the meta-language of the syntactic model referred to and seems to think of case distinctions. It is, namely, proposed a categorization such as Ergative and Locative for nouns within clauses of a certain kind, where Ergative replaces S, which would change the whole paradigm. Moreover, any syntactic model defines the linguistic meaning of nouns through the verb. Even if S and D get their syntactic markers in memory, it is unclear which ones and how, since the formula indicates no meaning for verbs and no rules are given. With such a non-linguistic treatment of the verb special problems are raised in explaining the M parameter, which is particularly syntactic.

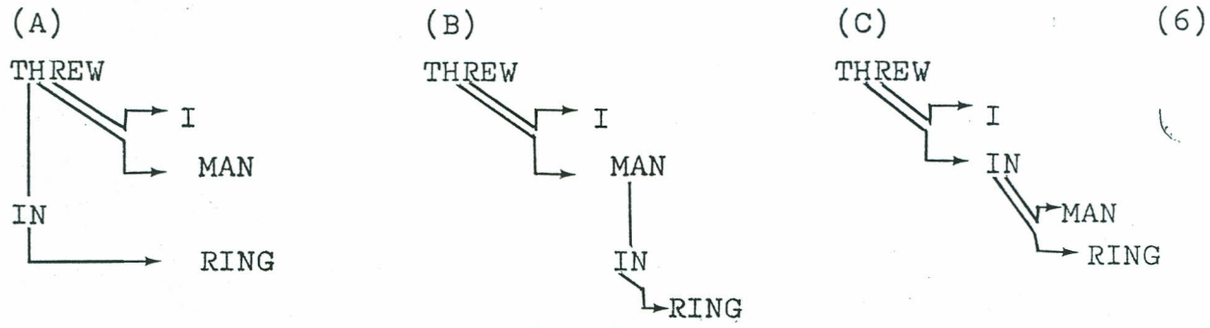
The selection process was manually performed by coders. Quillian argues that even if the coders are trained to decide the " exact " meaning of a sentence, they are " most unreliable and unhappy about making this distinction"(p 250-251). Thus Quillian discusses the need for a model by which ambiguity can be solved. The sentence

I threw the man in the ring (5)

is given three different meanings:

(a) While in the ring I threw the man
(b) I threw the man who was in the ring
(c) I threw the man into the ring

The encodings correspond to



(6)

The proposition is " I threw the man ", whose unambiguous coding is indicated by paranthesis. Regardless of whether the three meanings can be logically represented in his graph or not, Quillian decides for the following codings for the rest of the sentence (the relation of the proposition to " the ring "). Meanings of type (c) are coded like (A), and meanings like (a) are coded as modifiers of the subject. This eliminates forms of type (C) and brings the codes more in line with standard terminology, Quillian declares.

Even though meaning (a) could be inferred from the sentence (5), (A) is not the corresponding notation. The link indicates a modification of a relation, which seems unfounded. If, instead, it represents a modification to the entire clause, this means that the formula is ambiguous and that meaning (a) is not correctly understood. Otherwise it would not have been necessary to reformulate the coding instruction, and the formulation would have looked like (B), principally. The result is a fourth meaning. A new paranthesis notation is added to allow " in the ring " to modify the rest of the sentence. This is what is intended in (A). From a theoretical point of view, eliminating (C) in favour

of (A) means a destruction of the meanings pertaining to the verb array. Thus meaning (c), which is the only grammatically correct one in its sentence form, cannot be represented, and the prepositional phrases with directional and locational sense respectively are not distinguished. Paradoxically, most of the linguistic meaning that can be assigned to syntactically variable elements is subsumed under the M parameter.

Symbols and Contextual Constraints

A modelling of cognitive mechanisms, such as comprehension, based on natural language does not necessarily mean that language and cognition should be described in the same model. But, as Quillian has explicitly referred to Chomsky's syntactic model for extracting linguistic units meaningful for comprehension (concepts), he would be expected to demonstrate the representational link between syntactic components and the conceptual structure. The formula (4) clearly indicates a view of structure as propositional knowledge, a view common within wide circles. However, the formula differs from Chomsky's model in the distinction and specification of the symbols, which makes the reference to it unclear as to the kind of meaning to be represented, or the degree to which Chomskyan notions apply to the symbolic unit.

The reader of Quillian's discussion of meaning could easily realize his frustration about the inability of his propositional unit to take care of all syntactic information in his example sentence (5). He choses a " common sense " way out of the problem. Moreover, a TG description does not suggest a " memory ", since all experience is preserved in the formalism. It is very likely that Quillian's efforts to use this formalism for building a memory device is due to the fact that Chomsky was the only linguist at that time who claimed to have constructued a computationally well-suited formalism with an anchorage in theory. This theory is expressed in the naming of the symbols and their operation.

Structure and representation. The view of language as an abstract phenomenon, anchored in human biology, is stated in the Chomskyan symbols. In being abstract, language can only be studied through representation. The representation in the form of a grammar is, therefore, often taken for the structure itself. With this position Chomsky is able to state that deep in the human biology (structure) we find Noun Phrase and Verb Phrase conceived as categories accordingly. But categories emerge from an identification process (Bierschenk, 1984). With the mathematical-logical formalism as descriptive instrument they are, of course, artificial categories and should be termed " classes ".

A common conception among linguists is that the rules of grammar, syntactic organization in particular, are the result of conventions. To use these conventions in the set up of a computer model for describing the organization is a technical invention. Yet, there are computer oriented scientists who conceive it as a detection. To make it a theory, there has to be empirical evidence that the syntactic organization corresponds to structure. It would perhaps be fruitful to ask whether " natural grammar " really is convention.

Procedural descriptions. Theoretical linguists have concentrated on the definition of what language is. This definition has become obscured through the kind of formalism that defines how language is " recognized ", " generated ", or " comprehended ". Thus into the declarations are slipped elements belonging to processing. A contributing factor is Chomsky's concept of transformation, which introduces a procedure in the description. A key verb of procedure is " parse ", which means to break a sentence down into its component parts of speech with an explanation of the form, function, and syntactic relationship of each part. The procedural specification of a parser is commonly confused with perceptual ability and a cognitive representation strategy. Because, natural language comprehension is tested in

comparison with a parsing algorithm, based on artificially constructed sentences, for the most part. Therefore, to many linguists people parse sentences when they build a cognitive structure out of a text.

This element-based view of the relationship between linguistic units and comprehension can be seen, for example, in the ATN (Augmented Transition Network) grammars. To the description are added a hold or stack mechanism (short-term memory), a time schedule for the ordering of steps, and a heuristic filter for suggestion of structural descriptions. What motivates the execution order to be incorporated into the definition of language? (cf Dresher & Hornstein, 1976).

Reference Factors

The goal of a comprehension process is understanding, which presupposes a dynamic relationship between the comprehender and that which shall be comprehended. No comprehension can come about under static conditions. This means that viewpoints must vary or be viewed from different angles, and the viewer must be able to shift position, because his perspective(s) determines what is perceived and how it is comprehended. Identification of invariant structure constitutes the prerequisite for the " direct perception " of information from symbols that characterizes language comprehension processes. Then, what theoretical considerations underlie models of comprehension based on the perception of strings of linguistic elements?, or, in other words, where is the point of reference?

There is no question about the stationary relationship in a propositionally defined information unit. It symbolizes a universal meaning, which implies that perspective and viewpoints are freezed and can no longer be detected. This circumstance makes ambiguity grow and makes model constructors invent rules and regulations out of their own common sense, that is, they interpret, since there is no empirical context for direct comprehension. In this light, it is not difficult to realize that the

symbols of the meta-language are the meaning itself. So, the symbols are empty; language carries nothing at all. How can one possibly explain that so many efforts on simulating and testing natural language understanding are based on a " comprehensional zero hypothesis "?

References

Bierschenk, B. Steering mechanisms for knowability. Cognitive Science Research (Lund: Lund University), May, 1984.

Bobrow, D. & Winograd, T. An overview of KRL, a knowledge representation language. Journal of Cognitive Science, 1977, 1, 3-46.

Chomsky, N. Syntactic structures. The Hague: Mouton, 1957.

Chomsky, N. Aspects of the theory of syntax. Cambridge, Mass.: MIT Press, 1965.

Dresher, B.E.,& Hornstein, N. On some supposed contributions of artificial intelligence to the scientific study of language. Cognition, 1976, 4, 321-398.

Quillian, M.R. Semantic memory. In M. Minsky (Ed.), Semantic information processing. Cambridge, Mass.: Cambridge University Press, 1968. Pp. 216-270.